Noname manuscript No.
(will be inserted by the editor)

Comment on Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination

Stefan Van Aelst

Received: date / Accepted: date

Agostinelli, Leung, Yohai and Zamar (Agostinelli et al. in the remainder) consider the difficult problem of robust estimation based on high-dimensional data. If outlying values can appear independently in the variables, then it can easily occur that the majority of the observations in high-dimensional data is contaminated, as pointed out in Alqallaf et al (2009). Consequently, standard robust methods fail in this case, and new methods need to be developed that can handle this type of contamination. Moreover, next to independent contamination also casewise or structural outliers can still appear in the data. This situation was formalized as the partially spoiled independent contamination model (PSICM) in Alqallaf et al (2009).

In their paper Agostinelli et al. are the first to introduce a consistent estimator of multivariate location and scatter that is highly robust against both cellwise and casewise outliers. The 2SGS is a strongly consistent estimator of the location and shape of general elliptical distributions. Similarly to other proposals, the estimator proceeds in two steps. In the first step an outlier detection rule is used to identify potential cellwise outliers. A first improvement is the use of a data adaptive cutoff instead of a fixed cutoff value when filtering cellwise outliers. The second novelty is to replace flagged outliers by missing values as first proposed in Danilov (2010) and Farcomeni (2013), while earlier proposals tried to reduce their effect through some form of Winsorization, see e.g. Alqallaf et al. (2002), Van Aelst et al. (2011,2012), Van Aelst (2015). In the second step, the location and scatter are estimated based on the data set with missing values by using the GSE estimator of Danilov et al (2012). GSE

S. Van Aelst
KU Leuven, Department of Mathematics, Section of Statistics,
Celestijnenlaan 200B B-3001, Leuven, Belgium and
Ghent University, Department of Applied Mathematics, Computer Science and Statistics,
Krijgslaan 281 S9, Gent, Belgium
E-mail: Stefan.VanAelst@wis.kuleuven.be

is a very effective estimator, but is also computationally very demanding. This limits its use for really high-dimensional data sets, e.g. $p \geq 100$.

The replacement of cellwise outliers by missing values opens the door to apply missing data methods to the incomplete data. For example, instead of directly estimating the parameters from the incomplete data by a complex estimation procedure, an initial imputation step can be applied. If the filtering and imputation are successful, then the imputed data will only contain casewise outliers. Thus, any standard robust estimation method can be used to estimate the parameters from the imputed data. Hence, computationally efficient procedures such as the fast MCD (Rousseeuw and Van Driessen 1999) or the fast S/MM (Salibian-Barrera and Yohai 2006, Salibian-Barrera et al. 2006) can be used. For very large, high-dimensional data the recently developed deterministic MCD (Hubert et al. 2012) or S/MM (Hubert et al. 2015) can be used.

When imputing the data, we need to take into account that the inserted missing values are not missing completely at random. However, the missingness is non-informative in the sense that the recorded value was an outlying value which did not provide any useful information. Moreover, if we make the common assumption that the cellwise contamination indicator $\mathbf{B}_\epsilon$ in ICM (see expression (2) of Agostinelli et al.) is independent of both $\mathbf{X}_0$ and $\tilde{\mathbf{X}}$, then we can use the data distribution to impute the missing values. An overview of such imputation methods can be found in e.g. Cevallos Valdiviezo and Van Aelst (2015). However, since the incomplete data still may contain structural outliers, a robust imputation strategy should be used (see e.g. Vanden Branden and Verboven 2009).

To illustrate these ideas, I consider the following procedure.

Step I. *Eliminate cellwise outliers* by using the Gervini-Yohai filter and replace them by NA's.

Step II. *Impute the missing values.* To impute the missing values, I use the following simple procedure. For each empty cell, determine the most correlated variable with a non-empty cell for that same case. The correlation is measured robustly by using the Gnanadesikan and Kettenring (1972) procedure based on the efficient and robust Q_n estimator of scale (Rousseeuw and Croux 1993). Take this variable as the regressor in a robust simple regression that uses all the complete cases for the two variables. I used the MM-estimator of Yohai (1987) for this purpose. Assuming that the errors are normal, determine the predictive distribution for the empty cell and impute the cell by making a random draw from this distribution.

Step III. *Robustly estimate the location and scatter* from the imputed data. To see whether this procedure gives an improvement over the Huberized Stahel-Donoho (HSD) estimator which is considered in Agostinelli et al., I used the Stahel-Donoho (SD) estimator to estimate the parameters.

The imputation technique in Step II is a simple attempt to approximate the conditional distribution of the variable with missing value based on the available data. Of course, more complex methods can be developed to bet-

ter approximate this conditional distribution by using all the variables with an observed value for the case whose cell needs to be imputed. To keep the computation time low, only one imputation is drawn from the predictive distribution. However, it is straightforward to generate multiple imputed data sets from which the parameters can be estimated more precisely in Step III.

To examine the performance of this imputation approach, Figure 1 shows the results of a simulation with the same design as in Agostinelli et al. Results are shown for data with 10% of contamination. The plots on the left in Figure 1 show the average LRT distances in function of k for ICM outliers and can be compared to Figure 1 in Agostinelli et al. The plots on the right in Figure 1 show the average LRT distances in function of k for THCM outliers and can be compared to Figure 2 in Agostinelli et al. Next to the imputed data SD estimator (ISD) the results for the standard SD estimator are shown as well.

As could be expected, the ISD estimator performs not as good as the SD estimator for THCM contamination. For this model ISD shows a pattern of behavior that is similar to HSD in Figure 2 of Agostinelli et al., but with somewhat larger distances. However, in case of ICM contamination, ISD shows a behavior that is similar to the 2SGS estimator and thus behaves much better than HSD. In fact, it can be seen that for this type of data with a correlation matrix that has a high condition number, there is not much difference between the HSD and SD estimators. Hence, the Winsorization in HSD is not effective in this setting. Overall, these limited results suggest that ISD can handle both cellwise and casewise outliers. Three step estimators in which an initial filtering is followed by a suitable robust imputation procedure may be a viable alternative to robustly analyze high-dimensional data. Due to the flexibility to choose an appropriate robust estimation procedure in the third step, it may be easier to extend this approach to handle large data sets in really high-dimensions.

References

1. Cevallos Valdiviezo H, Van Aelst, S (2015) Tree-Based Prediction on Incomplete Data Using Imputation or Surrogate Decisions. *Information Sciences* 311: 163–181.doi: 10.1016/j.ins.2015.03.018
2. Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* 28: 81–124.
3. Hubert M, Rousseeuw PJ, Vanpaemel D, Verdonck T (2015) The DetS and DetMM estimators for multivariate location and scatter. *Comput. Statist. & Data Anal.* 81: 64–75.
4. Hubert M, Rousseeuw PJ, Verdonck T (2012) A deterministic algorithm for robust location and scatter. *J. Comput. & Graphical Statist.* 21: 618–637.
5. Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.* 88: 1273–1283.
6. Rousseeuw PJ, Van Driessen K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41: 212–223.
7. Salibian-Barrera M, Van Aelst S, Willems G (2006) PCA based on multivariate MM-estimators with fast and robust bootstrap. *J. Amer. Statist. Assoc.* 101: 1198–1211.
8. Salibian-Barrera M, Yohai V (2006) A fast algorithm for S-regression estimates. *J. Comput. & Graphical Statist.* 15: 414–427.

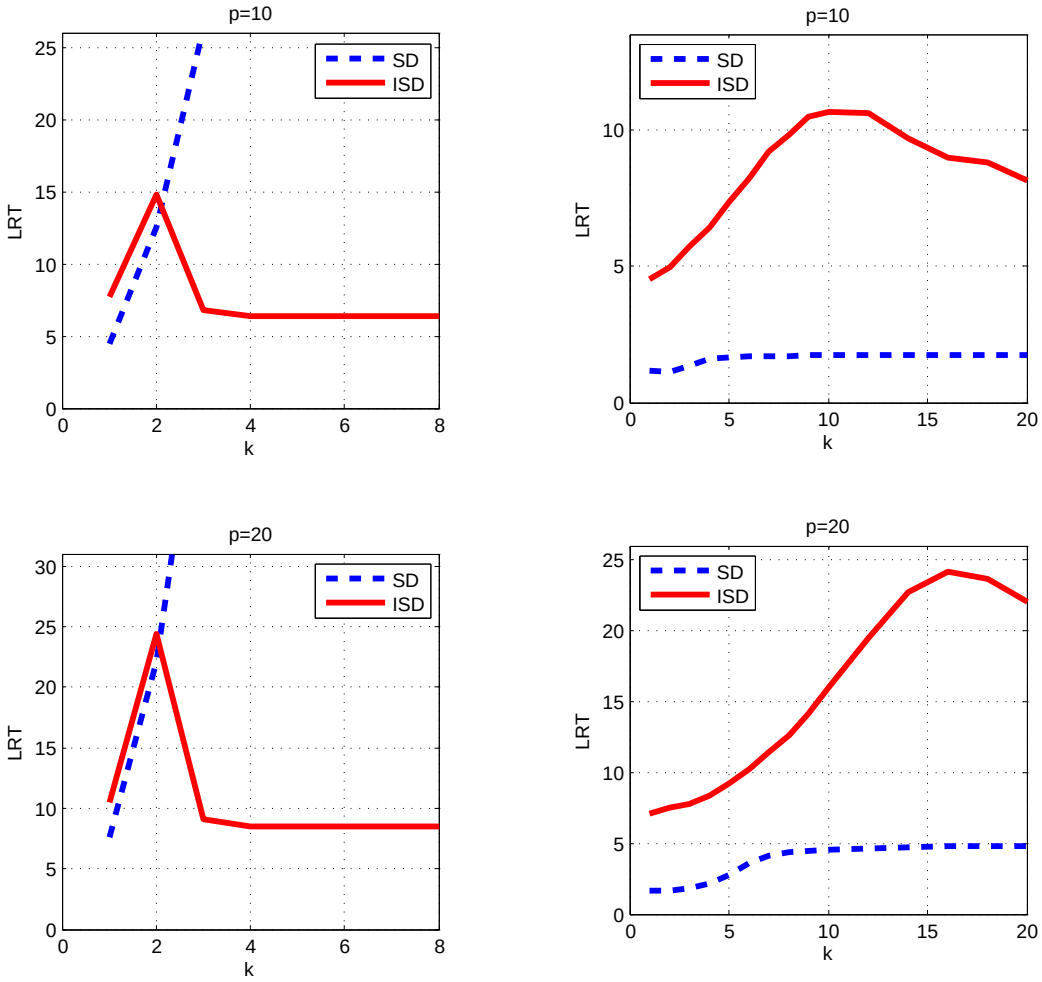

Fig. 1 Average LRT distances for 10% of contamination at different values of k. Left plots show results for ICM, right plots for THCM.

9. Van Aelst, S. (1015) Stahel-Donoho estimation for high-dimensional data. *Inter. J. Computer Math.*, to appear. doi: 10.1080/00207160.2014.933815
10. Van Aelst S, Vandervieren E, Willems G (2011) Stahel-Donoho estimators with cellwise weights. *J. Statist. Comput. and Simul.* 81: 1–27.
11. Vanden Branden K, Verboven S (2009) Robust data imputation. *Comput. Biology and Chemistry* 33, 7–13.
12. Yohai VJ (1987) High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* 15: 642–656.